Are Standard Diagnostic Test Characteristics Sufficient for the Assessment of Continual Patient Monitoring?

Liangyou Chen, PhD, Andrew T. Reisner, MD, Xiaoxiao Chen, PhD, Andrei Gribok, PhD, Jaques Reifman, PhD

Background. *For diagnostic processes involving continual measurements from a single patient, conventional test characteristics, such as sensitivity and specificity, do not consider decision consistency, which might be a distinct, clinically relevant test characteristic.* ***Objective****. The authors investigated the performance of a decision-support classifier for the diagnosis of traumatic injury with blood loss, implemented with three different data-processing methods. For each method, they computed standard diagnostic test characteristics and novel metrics related to decision consistency and latency.* ***Setting.*** *Prehospital air ambulance transport.* ***Patients.*** *A total of 557 trauma patients.* ***Design****. Continually monitored vital-sign data from 279 patients (50%) were randomly selected for classifier development, and the remaining were used for testing. Three data-processing methods were evaluated over 16 min of patient monitoring: a 2-min moving window, time averaging, and postprocessing with the sequential probability ratio test (SPRT).* ***Measurements****. Sensitivity and specificity were computed. Consistency was quantified through cumulative counts of decision changes over time and the fraction of patients affected by false alarms. Latency was evaluated by the fraction of patients without a decision.* ***Results****. All 3 methods showed very similar final sensitivities and specificities. Yet, there were significant differences in terms of the fraction of patients affected by false alarms, decision changes through time, and latency. For instance, use of the SPRT led to a 75% reduction in the number of decision changes and a 36% reduction in the number of patients affected by false alarms, at the expense of 3% unresolved final decisions.* ***Conclusion****. The proposed metrics of decision consistency and decision latency provided additional information beyond what could be obtained from test sensitivity and specificity and are likely to be clinically relevant in some applications involving temporal decision making.* ***Key words:*** *continual patient monitoring; decision-support algorithm; sequential probability ratio test; physiological data.* ***(Med Decis Making 2013;33:225–234)***

Continual physiological monitoring is standard practice in many health care arenas, e.g., hospital wards and operating rooms, where vital-sign data are measured repeatedly so that if instability occurs it can be detected and treated promptly. However, false alarms are a major problem because standard alarms are triggered when certain parameter thresholds are reached.[1–3] All too often, the abnormality that triggers an alarm is either a measurement artifact or a benign transient event. Yet, when false alarms occur frequently, there is a deleterious effect on patients in that caregivers may be slow to respond to alarms with low positive predictive value.[4]

Received 10 August 2011 from DoD Biotechnology HPC Software Applications Institute (BHSAI), Telemedicine and Advanced Technology Research Center (TATRC), Fort Detrick, MD (LC, ATR, XC, AG, JR), and Massachusetts General Hospital Department of Emergency Medicine, Boston, MA (ATR). This work was performed at BHSAI, TATRC, US Army Medical Research and Materiel Command (USAMRMC), Fort Detrick, MD 21702. This work was partially supported by the Combat Casualty Care Research Area Directorate of the USAMRMC, Fort Detrick, MD. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The following author is employed by the sponsor: Jaques Reifman is a U.S. Department of the Army Senior Research Scientist. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense. This article has been approved for public release with unlimited distribution. Revision accepted for publication 24 March 2012.

DOI: 10.1177/0272989X12451059

Address correspondence to Jaques Reifman, BHSAI, TATRC, ATTN: MCMR-TT, 504 Scott Street, Fort Detrick, MD 21702-5012; e-mail: jaques.reifman@us.army.mil.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
FEB 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Are Standard Diagnostic Test Characteristics Sufficient for the Assessment of Continual Patient Monitoring?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Medical Research and Materiel Command,DoD Biotechnology High Performance Computing Software Applications Institute,Telemedicine and Advanced Technology Research Center,Fort Detrick,MD,21702**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**10**	

In this report, we considered a set of analytic methods for detecting abnormalities from continual physiological data and examined how the techniques compared through time. We examined whether standard test characteristics (sensitivity and specificity) were adequate for describing the resultant alarm behaviors from one time interval to the next. Specifically, we developed metrics to measure the temporal stability of test decisions, which we termed *consistency,* and examined the extent to which patient alarms were consistent through time. Our intent was to describe whether alarms tended to reoccur in the same patients from one time period to the next (on whom the clinical staff would be able to focus attention) or if (false) alarms were distributed throughout the entire monitored population (so that many disparate patients would—unnecessarily—require attention as the alarms were triggered).

We focused on several basic methods for pre- and postprocessing of continual vital-sign data into and out of a core alarm algorithm. Analytic methods for identifying irregularities from a set of time-series data have been well established in the manufacturing quality control literature. Methods dealing with this problem include the sequential probability ratio test (SPRT),[5,6] which evaluates the likelihood ratio of 2 hypotheses based on sequentially available evidences. Alternatives include the control chart method,[7,8] the belief-modeling method,[9] and other Bayesian-based methods.[10,11] Among these methods, the SPRT is simple to calculate and, for given false-positive and false-negative probabilities, requires the smallest number of samples to achieve a decision (unless the statistical model is grossly incorrect).[5]

Our goal was to investigate if conventional test characteristics were adequate for assessing the basic performance of an alarm or if it was also necessary to consider its temporal consistency. In a comparative analysis, we employed 3 methods for pre- and postprocessing of continual data into and out of our core alarm algorithm. Compared with a 2-min moving window, we examined if additional time averaging and the SPRT could alter the overall accuracy, the temporal consistency, and the latency of the algorithm output. The core alarm algorithm was a multivariate classifier for the diagnosis of traumatic injury with blood loss, given data from a standard prehospital patient monitor.[12] This analysis has implications for any diagnostic process involving continual vital-sign measurements from a single patient.

METHODS

This is a retrospective, comparative analysis, based on a previously reported ensemble classifier,[12] which provides automated detection of traumatic injury with blood loss in prehospital patients based on basic vital signs. We simulated 3 methods to process real-time data during the initial 16 min of prehospital patient transportation. The moving window method involved a moving 2-min analysis window; at every moment in time, the classifier was applied to the most recent 2 min of physiological data. The time-averaging method analyzed all available data from a given patient, from the onset of the data record to the current time (up to a maximum of 16 min). In the SPRT method, we applied the SPRT to the output of the classifier.

Trauma Patient Data

The physiological time-series data were collected from 643 trauma-injured patients during their first 16 min of helicopter transport to a trauma center.[13] The time-series variables were measured by ProPaq 206EL vital-sign monitors (Protocol Systems, Beaverton, OR) and consisted of electrocardiogram, photoplethysmogram, and respiratory waveform signals recorded at various frequencies and their corresponding monitor-calculated vital signs, including heart rate (HR), respiratory rate (RR), and arterial oxygen saturation (SaO_2), recorded at 1-s intervals, and systolic (SBP), mean, and diastolic (DBP) blood pressures collected intermittently at multiminute intervals.

We performed chart reviews to determine whether the transported trauma patients had traumatic injury with blood loss. Traumatic injury with blood loss was defined as requirement of a blood transfusion within 24 h upon arrival at the trauma center and also documentation of an explicitly hemorrhagic injury, either a) laceration of solid organs, b) thoracic or abdominal hematomas, c) explicit vascular injury and operative repair, or d) limb amputation. Patients who received blood but did not meet the documented injury criteria (60 cases), and patients who died before arrival at the hospital (26 cases) were excluded from the analyses because of uncertainty about whether they truly suffered traumatic injury with blood loss. Thus, we used a total of 557 patients, of which 61 were categorized as patients with traumatic injury and blood loss and the remaining 496 as controls.

Decision-Support Classifier: Training

The ensemble classifier aggregates 25 least-squares linear classifiers, each trained with a different subset of 5 input variables (HR, RR, SaO_2, SBP, and DBP) and with target values of 0.0 and 1.0, standing for control and traumatic injury with blood loss outcomes, respectively, to generate an (arithmetic) average output that can be used to discriminate the 2 outcomes.[12] We assigned ensemble-averaged outputs of ≤ 0.5 as control outcomes and outputs of > 0.5 as traumatic injury with blood loss. The ensemble classifier has been shown to provide more consistent performance than a single linear classifier, and importantly, it accommodates missing data, providing an output as long as any 1 of the 5 inputs is available.[12]

We randomly selected 50% of the study population (279 patients; 248 controls and 31 patients with traumatic injury and blood loss) to train (i.e., develop) the classifier. In prior studies,[14] we found that prehospital vital-sign data are very noisy, and hence, we developed algorithms that automatically assess the reliability of each vital sign used as input to the classifier.[15–17] We also reported that reliable data are diagnostically superior to unreliable data.[15,18] In another study,[14] we found that there are no major time-series trends in these vital-sign data, and averaging the most reliable data during transport yielded the best discriminatory performance. Consequently, we used the average value of the most reliable training data points from the first 16 min of transport time as input to train the ensemble classifier.

Evaluation of the Moving Window, Time-Averaging, and SPRT Methods

We investigated 3 methods to pre- and postprocess the ensemble classifier data. In each method, 1) the first 2 min of transport vital-sign data were used as a buffer where no classifications were made; 2) every 1 s we averaged the most reliable available vital-sign data (HR, RR, etc.) over a specified time window, input the averaged values to the classifier, and obtained an output; and 3) every 15 s, we averaged the previous 15 classifier outputs to generate a decision. The 3 methods differed on the length of the preprocessing time window of the classifier input data in item 2 (above) and on any additional postprocessing in the classifier outputs in item 3.

For the moving window, we averaged the classifier inputs over a 2-min time window and compared the averaged decision every 15 s with a 0.5 threshold. The time-averaging method differed from the first method in that the length of the time window for averaging the vital-sign input data grew continually up to the current decision time so that all available data were considered for each decision. In the SPRT method, the classifier outputs were further processed as inputs to the SPRT to generate a SPRT decision (or no decision), as described below.

The Sequential Probability Ratio Test

We investigated the ability of Wald's SPRT[5,6] to consider the sequential nature and postprocess the outputs of the ensemble classifier while balancing decision accuracy, consistency, and latency. Given a sequence of outputs $Y_1, Y_2, \ldots$ not necessarily independent from the ensemble classifier, so that $Y = N(\mu_Y, \sigma_Y^2)$ is a normal Gaussian process with an unknown mean μ_Y and a known constant variance σ_Y^2, the SPRT classifies a patient as control or traumatic injury with blood loss, or makes no decision, based on hypothesis testing. Note that σ_Y^2 was estimated as the variance of the ensemble classifier outputs at the end of the transport, i.e., at 16 min, and was kept fixed throughout the analysis. The SPRT tests a null hypothesis (H_0) that $\mu_Y = \mu_0$ against an alternative hypothesis (H_1) that $\mu_Y = \mu_1$, where μ_0 and μ_1 denote the arithmetic mean values of the classifier outputs for the control and traumatic injury with blood loss cases, respectively, with $\mu_0 < \mu_1$. If we let p_0 and p_1 be the probability density functions governing the two hypotheses, H_0 and H_1, respectively, then the observed likelihood ratio at decision time J can be represented as $l_J = \prod_{j=1}^{J} \frac{p_1(Y_j)}{p_0(Y_j)}$, with $J = 1, 2, \ldots$.

According to Wald's SPRT methodology,[5] we

$$\begin{gathered}\text{accept } H_0 \text{ (control)}, \text{if } \log(l_J) < B; \text{or} \\ \text{accept } H_1 \text{ (traumatic injury with blood loss)}, \\ \text{if } \log(l_J) > A; \text{or} \\ \text{continue to decision time } J+1, \text{if } B \leq \log(l_J) \leq A,\end{gathered} \quad (1)$$

where A and B are constants, with $0 < B < A < \infty$, chosen using Wald's criteria,[5] as to yield nominal false-positive probability (α; $0.0 < \alpha < 0.5$) and nominal false-negative probability (β; $0.0 < \beta < 0.5$) as follows:

$$\begin{aligned} A &= \log\frac{1-\beta}{\alpha}, \text{ and} \\ B &= \log\frac{\beta}{1-\alpha}. \end{aligned} \quad (2)$$

When α and β are relatively small (e.g., < 0.05), the SPRT tends to delay making a decision until

additional corroborating classifier outputs become available. Conversely, when α and β are large (e.g., ≈ 0.5), the SPRT makes quicker, albeit less accurate, decisions. Thus, by appropriately selecting these two parameters, we can balance decision accuracy, consistency, and delay. To this end, we determined the nominal probabilities α and β by minimizing a cost function ¢, which linearly combined the accuracy of the decisions, defined by its sensitivity (S_e) and specificity (S_p), at the end of the transport (i.e., at 16 min); the cumulative incidences of decision changes (D_c; from control to traumatic injury with blood loss and vice versa) over the 16 min of transport time; and the fraction of patients with no decision (N_d) at the end of the transport. Accordingly, we defined ¢ as follows:

$$¢ = \frac{1-S_e}{0.05} + \frac{1-S_p}{0.05} + \frac{D_c}{10} + \frac{N_d}{0.01}, \quad (3)$$

where the rescaling factors of the summands were empirically obtained through SPRT trial simulations on the training data so to normalize the effect of each of the four summands on ¢.

Under the Gaussian model, the log-likelihood ratio $\log(l_J)$ in equation 1 can be recursively calculated as follows:

$$\log(l_{J+1}) = \log(l_J) + \frac{\mu_1 - \mu_0}{{\sigma_Y}^2}\left(Y_{J+1} - \frac{\mu_1 + \mu_0}{2}\right), \; J = 0, 1, 2, \ldots, \quad (4)$$

where the initial log-likelihood $\log(l_0)$ can be selected arbitrarily and was set to 0.0 in this study. While it has been shown that the SPRT achieves a selected confidence in the shortest decision time,[5] it may not always arrive at a decision. However, when a decision was made, we noted the decision, stuck to it, and restarted the SPRT process from that time point until a new decision emerged.

Investigational Metrics

We compared the performance of the 3 data-processing methods using testing data from 278 patients where we evaluated the accuracy, latency, and consistency (in a sense to be defined) of the methods in aggregate using the following 5 performance metrics:

1. Sensitivity: at any given time *t*, the fraction of patients with traumatic injury and blood loss who were correctly identified by the algorithm at time *t*;
2. Specificity: at any given time *t*, the fraction of control patients who were correctly identified by the algorithm at time *t*;
3. No decisions: at any given time *t*, the fraction of patients without a decision out of the total number of patients;
4. Cumulative decision changes: the cumulative count up through time *t* of decision changes D_c; and
5. False-alarm-affected patients: the fraction of control patients incorrectly identified as having traumatic injury with blood loss, at or before time *t*, out of the total number of patients.

Every 2 min, from 2 to 16 min of transport time, we performed statistical tests of significance with pairwise comparisons between the investigational methods (i.e., moving window, time averaging, and SPRT). For proportions (sensitivity, specificity, no decisions, and false-alarm-affected patients), we employed Liddell's exact test.[19] The counts of total decision changes throughout the population cannot be statistically evaluated, so we also computed the total decision changes *per subject* and applied the Wilcoxon signed–rank test to the distributions. For all statistical tests, we considered a P value of < 0.05 to be statistically significant.

RESULTS

Figure 1 illustrates the continual output of the 3 data-processing methods, the moving window, time-averaging, and SPRT methods, in monitoring 4 control subjects (panel A) and 3 subjects with traumatic injury and blood loss (panel B). Each tile in the figure represents a 15-s outcome decision, with red (or dark) representing traumatic injury with blood loss decisions, green (or medium gray) control, and yellow (or light gray) no decisions. The selected control subjects illustrate different patterns in outcome decisions that we observed in the 248 control subjects in the testing data. For example, for subject 364, all 3 methods made correct and consistent control decisions over the 16-min transport time. For subject 607, each method generated some false-positive (i.e., false traumatic injury with blood loss) decisions. However, the moving window method generated the most frequent number of decision changes (3 changes from control to traumatic injury with blood loss and 3 from traumatic injury with blood loss to control, for a total of 6 decision changes), while the other 2 methods generated 2 decision changes each. For the third subject (640), unlike the other 2 methods, the SPRT method avoided making incorrect decisions (and decision changes), but the decision was delayed by more than 4 min. Finally, for subject 749, the SPRT was not able to make a definite decision during the

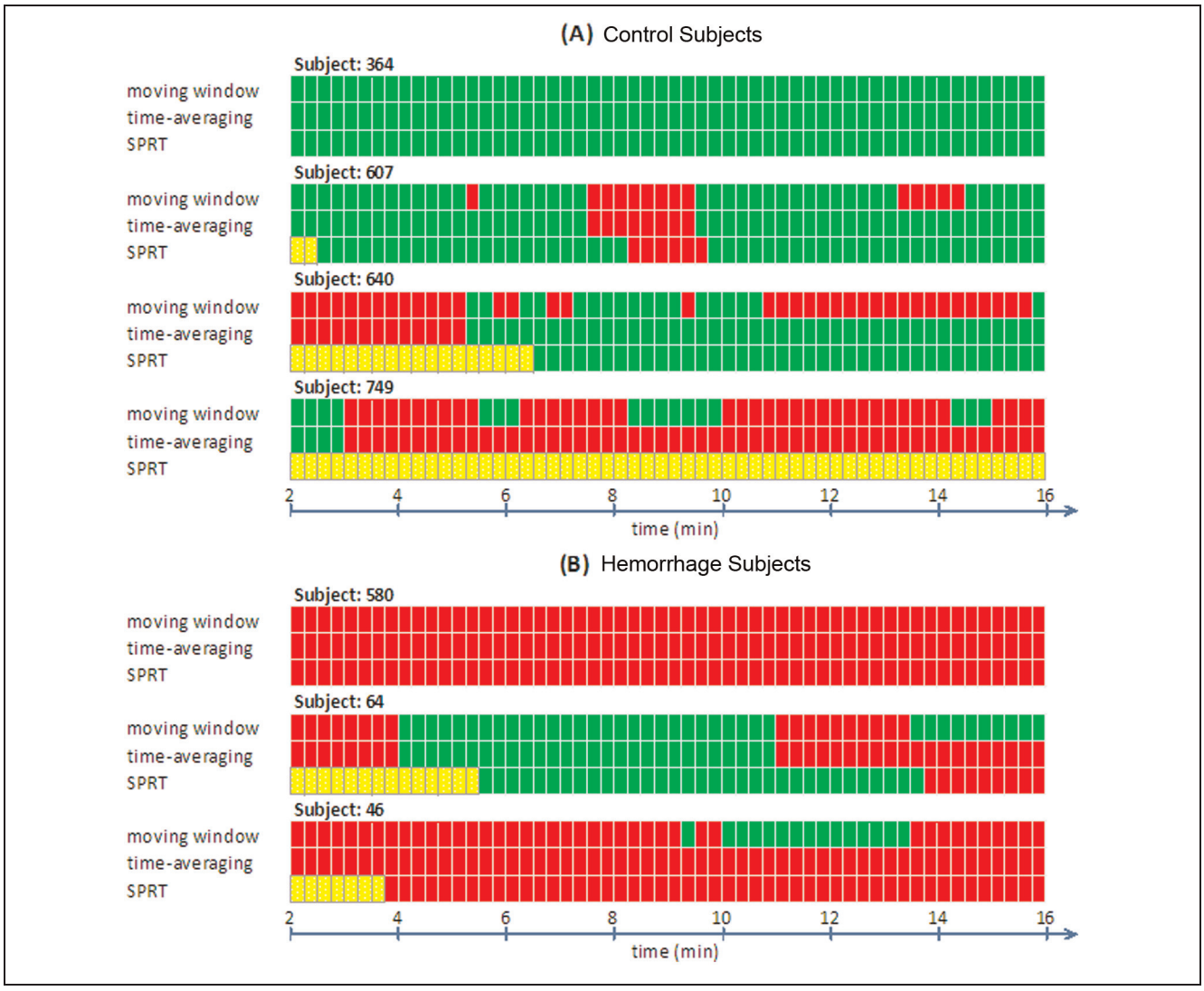

Figure 1 Continual outcome decisions over the 16 min of transport time for each of the 3 data-processing methods. (A) Selected pattern for 4 control subjects, and (B) 3 subjects with traumatic injury and blood loss. Each tile represents a 15-s outcome decision, with red (or dark) representing traumatic injury with blood loss decisions, green (or medium gray) control, and yellow (or light gray) no decisions. SPRT, sequential probability ratio test.

16-min transport time, while the other 2 methods generated decision changes and mostly incorrect decisions.

Panel B illustrates 3 patterns of decisions observed within the 31 patients in the testing set with traumatic injury and blood loss: for subject 580, all methods generated a consistent decision; for subject 64, the methods generated intermittent false-negative (i.e., false control) decisions, with the moving window method yielding an incorrect decision at 16 min; and for subject 46, all methods generated the correct final decision—however, the moving window produced decision changes and some incorrect decisions, while the SPRT did not produce a decision until almost 4 min.

Figure 2 illustrates the performance of the methods based on the 5 metrics (sensitivity, specificity, no decisions, cumulative decision changes, and false-alarm-affected patients) used to evaluate the accuracies, latencies, and consistencies of the methods for the 278 testing subjects over the 16-min transport time. Each of the 3 methods—moving window, time averaging, and SPRT—yielded comparable performance in terms of sensitivity and specificity at the end of the

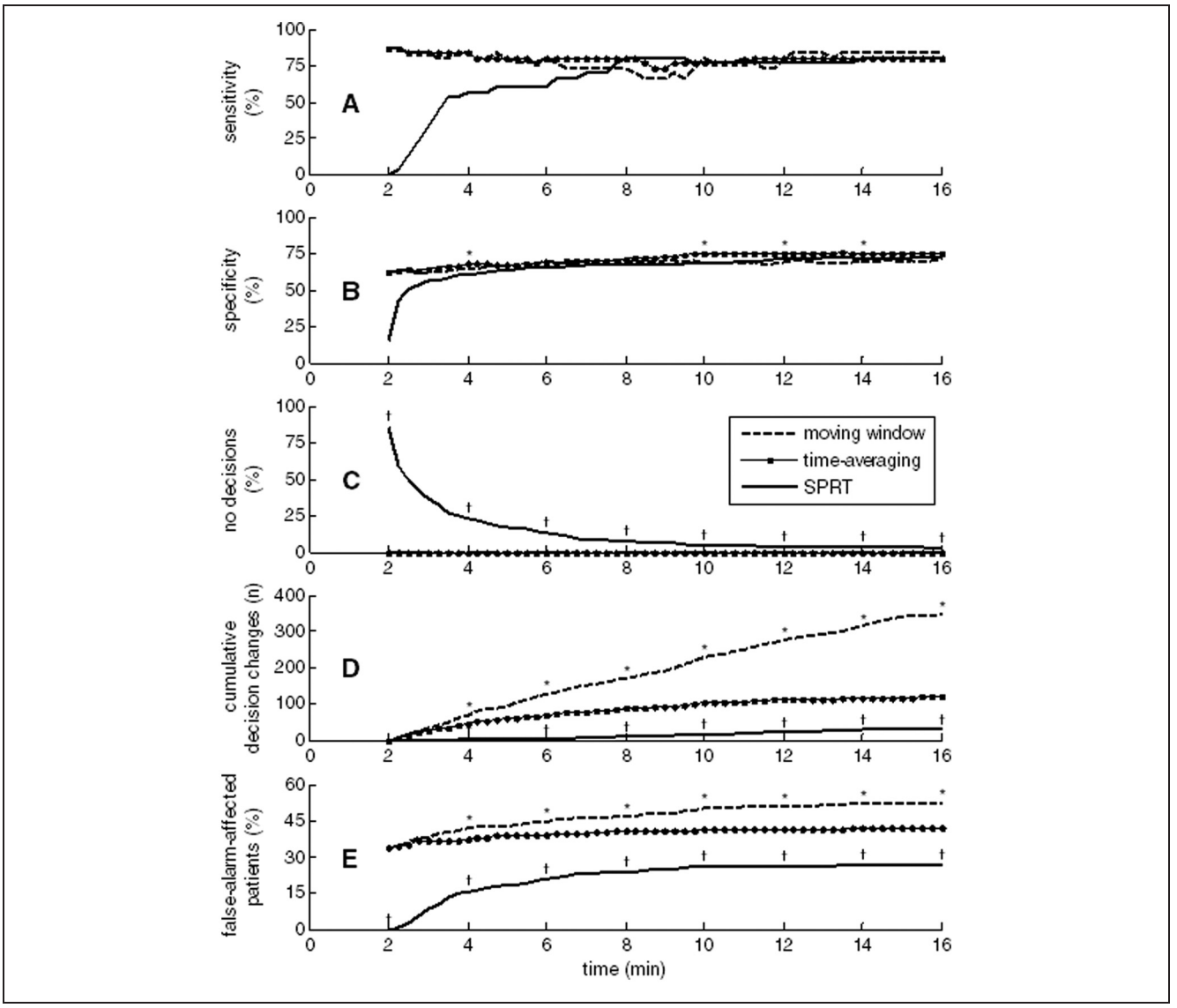

Figure 2 Comparison of 3 data-processing decision methods for the 278 testing subjects analyzed over the 16-min transport time based on 5 performance metrics: (A) sensitivity, (B) specificity, (C) fraction of patients with no decisions, (D) cumulative number of decision changes, and (E) false-alarm-affected patients. Pairwise tests of significance were performed every 2 min. Proportions were compared by Liddell's exact test (panels A–C, E). Panel D illustrates cumulative count of total-population decision changes, and the Wilcoxon signed–rank test was applied to the per patient counts of decision changes.$^{*}P < 0.05$, *time averaging v. moving window.* $^{\dagger}P < 0.05$, *SPRT v. both moving window and time averaging.*

transport time (sensitivity: 83%, 80%, and 80%, respectively; specificity: 71%, 75%, and 73%, respectively). Note that the SPRT method provided relatively low sensitivity and specificity ($\leq 60\%$) during the first 6 min of transport because of a large fraction of patients without SPRT decisions (see panel C). For instance, at 2 min, fewer than 25% of the patients had a decision rendered by the SPRT, and consequently, the corresponding sensitivity was also less than 25%. The SPRT method failed to make a decision at 16 min for 8 subjects (or 3% of the subjects), while the other 2 methods showed no decision latency (panel C).

In terms of consistency of decisions, the SPRT demonstrated a significantly reduced fraction of

false-alarm-affected patients throughout and at the end of the 16-min transport, compared with both other methods—27% of the subjects, which was 36% lower than the time-averaging method (42% of the subjects) and 48% lower than the moving window method (52% of the subjects; panel E). The SPRT also consistently generated fewer numbers of decision changes over time (29 total decision changes v. 118 for the time-averaging method and 348 for the moving window method; panel D).

The time-averaging method was more consistent than the moving window method, with significantly fewer false-alarm-affected patients and average decision changes *per patient*. The time-averaging method did not demonstrate the latency of the SPRT method.

DISCUSSION

In this article, we studied the accuracy, consistency, and latency of a decision-support classifier employing three different data-processing methods for the continual prehospital diagnosis of traumatic injury with blood loss in 557 trauma patients. It is striking that all methods showed very similar sensitivities and specificities yet very different temporal behaviors. For instance, Wald's SPRT was much more consistent, generating false alarms in significantly fewer patients, with significantly fewer decision changes.

There are 2 major implications. First, for some continual monitoring applications, standard test characteristics, e.g., sensitivity and specificity, are insufficient for describing the performance of a classifier because they do not describe if false alarms occur repeatedly in a limited subpopulation or if false alarms are evenly distributed throughout a population. Second, as a corollary, it is apparent that pre- and postprocessing of time-series data can significantly alter temporal consistency, as was seen in the application of time averaging and of the SPRT, a classic technique intended for precisely this type of application.

Insufficiency of standard test characteristics for describing the continual performance of a classifier. For the continual monitoring of patients, standard test characteristics do not consider the sequential nature of the algorithm's decisions when there are repeated decisions being made on each subject. For example, while 2 binary decision classifiers may have similar overall sensitivity and specificity, 1 may be less stable than the other, continually ''flipping'' its decisions through time (which is naturally exacerbated the more that a classifier is sensitive to transient noise in the signal). We found this exact phenomenon in our data set: After 5 to 10 min, the 3 investigational methods had similar sensitivities and specificities, but there were significant differences in the total number of patients affected by a false alarm. Using the SPRT significantly reduced the fraction of false-alarm-affected patients by approximately half, compared to the moving window method.

We speculate that this effect was notable in this analysis because the prehospital vital signs showed considerable intrasubject variability through time, with sizable fluctuations in HR, blood pressure, etc., during the course of prehospital transport.[14] Comparable fluctuations in the prehospital vital signs of trauma patients have been observed in other prehospital studies as well,[20–22] which may be physiological responses to episodic stimuli (e.g., pain and fear), to episodic therapies (e.g., fluids), or to underlying pathology, as well as some degree of routine biological variability and measurement error.

In general, are standard diagnostic test characteristics sufficient for the assessment of continual patient monitoring, or is it appropriate to quantify classifier consistency? It is likely that the frequency of decision changes in diagnostic classification is dependent on the classifier evaluation frequency, the temporal fluctuations in the diagnostic data, and the proximity of the classifier output to the decision boundary. Presumably, there is a continuum of diagnostic applications in terms of the classifier consistency through time. If the diagnostic data are temporally stable during intervals of disease and health, then standard test characteristics are likely sufficient. At the other extreme, if the diagnostic data fluctuate through time, then the diagnostic classification will also fluctuate through time, and it may be illuminating to consider metrics of consistency (as we have done in this report) in addition to standard test characteristics. In many reports, continual classifiers are evaluated without explicit consideration of their performance and consistency through time, such as reports by our group[12] and by others.[23–25] It is likely that, at least for a subset of continual monitoring applications, standard diagnostic test characteristics are insufficient and it would be valuable to consider consistency to quantify clinically relevant properties of the diagnostic test.

In addition, evaluating a temporal classifier through time can reveal if performance changes because of temporal disease progression. Presumably, it is easier to diagnose blood loss or septic shock as the pathology progresses, due to the spectrum

effect (e.g., when a diagnostic test performs better in a study population with more severe disease. Consider that the sensitivity of a hypothetical dip-test for leukocyte esterase in the diagnosis of urinary tract infection may be higher in patients of an underserved population, who tend to receive evaluation later in the course of disease, rather than in patients of an affluent population, who are promptly evaluated after the earliest symptoms). Spectrum effects also affect the temporal consistency of a diagnostic classifier, because small fluctuations in diagnostic data for a borderline case would be more likely to affect diagnostic classification (e.g., during early stages of blood loss). By contrast, cases with more advanced pathology will often have more frankly abnormal diagnostic data, and so temporal fluctuations are unlikely to alter diagnostic classification. That diagnostic classification may become easier as the disease process progresses is often well recognized. For instance, Cuthbertson[26] reported test characteristics for an investigative early warning score over hourly intervals, e.g., 1 h prior to patient acute deterioration, 2 h prior, etc. However, it was not reported to what extent the true and false alarms occurred in the same patients hour by hour, i.e., consistency. In this report, we describe the minute-by-minute performance of an investigational algorithm during the initial 16 min of prehospital transportation, including the temporal variation of decision changes in the same patients and the fraction of total patients affected by some of these changes. At least in our application, the additional statistics provide information beyond standard test characteristics, perhaps in part because we examined data measured soon after traumatic injury.

Pre- and postprocessing of time series alters performance of an automated continual classifier. Pre- and postprocessing of time-series data is appropriate for removing noise that occurs over faster time scales than the process of interest, thus enhancing the underlying signal. In this study, the narrow 2-min moving window caused a large number of patients to trigger false alarms (24% more than the time-averaging approach and 93% more than the SPRT approach). Failure of developers of monitoring algorithms to explicitly consider classifier output stability, or consistency, through time will presumably exacerbate the well-described problem of false alarms in medical monitoring systems[1–4] and will likely decrease the incentive for caregivers to adopt novel decision-support technologies. Conversely, excessively stable classifiers are also problematic, causing unacceptable latency when a patient's state does change. The challenge is to optimize the tradeoffs between classifier accuracy, consistency, and latency.

Consider time averaging. As long as the noise in the time series has no major bias, this is a practical technique for filtering out measurement error and transient physiological events. For a monitoring algorithm, the time-averaging window should be shorter than the onset time of the disease of interest. In other words, time averaging over 15 min may be useful when seeking hemorrhage physiology, although time averaging over 60 min might be too large a window, causing unacceptable latency to the detection of hemorrhage physiology that can progress in less than an hour. In this report, the time-averaging method was able to improve decision consistency (with 66% fewer decision changes) and reduce false-alarm-affected patients (with 20% fewer false-alarm-affected patients) compared with the simple 2-min moving window method.

A prior report corroborates this principle: that it is often possible to reduce false alarms at the expense of clinically acceptable latency. In monitoring children at home by pulse oximetry, Gelinas and others[27] suggested that the rate of hypoxia alarms ($SpO2 < 85\%$) could be reduced from 3.6 to 0.2 alarms per night without missing any clinically significant events, simply by requiring a 10-s duration of hypoxia (rather than alarming the instant that the hypoxia threshold was met).

The SPRT: a classic technique that can improve temporal consistency during continual monitoring. One classic application of the SPRT is for the evaluation of a shipment of manufactured components. Components are measured 1 by 1 until a SPRT decision is rendered that the set of components is within (or outside of) the acceptable tolerances. Our investigational algorithm is analogous in that measurements were taken repeatedly from 1 trauma patient, and the SPRT was used to decide whether the patient was within (or outside of) the range of vital signs typical of patients with traumatic injury and blood loss. Of course, given a shipment of components, individual measurements are statistically independent, while there is temporal correlation when measurements are repeated in the same patient. Regardless, our findings suggest that the SPRT is suitable for improving the consistency of the investigational classifier based on continual physiological data.

In the medical area, the SPRT has been previously applied to the performance monitoring of clinical teams[26–30] (to continually monitor the surgical outcome rate and ensure it does not deviate from the expected success rate), routine surveillance of drug

safety[31] (to continually monitor whether a new vaccine is safe over a period of time), and determination of early stopping criteria of clinical trials[32,33] (to allow the trial to be stopped as soon as the information accumulated is considered sufficient to reach a conclusion). Our results demonstrated that the SPRT may be effective for continual physiological monitoring, in the reduction of false-alarm-affected patients (36% fewer patients than the time-averaging method) and overall decision changes (75% fewer decision changes). The tradeoff was the occurrence of some decision latency because, unlike the other investigative methods, the SPRT can yield an ''undecided'' output (see Figure 2). Indeed, for several cases (3% of the total), there was never a diagnostic decision generated when applying the SPRT. For applications in which such a tradeoff is acceptable, the SPRT is optimal in the sense that, mathematically, it guarantees the smallest number of samples to achieve a decision for given false-positive and false-negative probabilities.[5] The performance of the SPRT depends on the selected nominal probabilities α and β, which can be set either arbitrarily or by optimizing certain cost function during classifier training. Properly chosen α and β may improve the sensitivity and specificity, and decrease the cumulative incidences of decision changes, with acceptable final unresolved decisions. However, improperly chosen α and β may significantly downgrade the sensitivity or the specificity. As well, when we first attempted to optimize the SPRT with a cost function customized wholly to yield small false-positive α and false-negative β probabilities, we improved the final accuracy but simultaneously increased the unresolved decisions to 40% on the testing data. In the end, the cost function defined in equation 3 provided a simple yet effective tool to balance accuracy, consistency, and latency.

This tradeoff between latency and consistency may limit the application of the SPRT in the detection of conditions that involve an imminent threat to life, e.g., cardiac tachyarrhythmia. However, in the monitoring of early disease states, when some latency is acceptable, e.g., early hemorrhage detection,[12] sepsis detection,[25,34] or other early warning functionality,[23,24] we suggest that the SPRT may provide a means to improve classifier stability and to reduce false alarms, without any necessary loss in decision accuracy.

Identification of traumatic injury with blood loss via continual physiological monitoring. The potential usefulness of the diagnostic classifier described in this report is not the focus of this study, and an assessment of potential clinical value must be tempered by the fact that the analysis is retrospective, based on post hoc classification as to whether each subject had traumatic injury with blood loss. Having said that, we believe that there is potential clinical value to the methodological application of conventional and commonsense analysis techniques to standard vital-sign data, e.g., noise rejection, time averaging, and multivariate classification. We previously found that automated techniques are diagnostically equivalent to prehospital severity scores based on medics' documentation.[15] In this case, we focused on the identification of hemorrhage because blood loss is 1 of the 2 primary reasons why trauma patients die,[35,36] but in many cases it can be treated effectively with blood transfusion and surgical hemorrhage control. We speculate that formal quantitative analysis of continual vital signs may be able to supplement today's convention, which relies on informal clinician judgments to integrate vital-sign data with other important clinical data. For instance, automated algorithms during prehospital care could be useful for triage and to aid the receiving hospital to efficiently mobilize proper resources, such as surgical teams and units of blood. Similar techniques could identify hospitalized patients who suffer unexpected episodes of blood loss during convalescence, e.g., early warning systems. However, actual performance and clinical usefulness must be prospectively assessed, and the optimal approach to decision support for trauma patients (e.g., attempt to identify any patients with traumatic injury and blood loss v. attempt to identify patients with uncontrolled, ongoing blood loss) involves open questions that are not addressed in this analysis.

CONCLUSION

Over time, all 3 methods converged to demonstrate very similar diagnostic accuracy (i.e., sensitivity and specificity). However, their consistency was significantly different. The SPRT significantly reduced the total number of patients affected by false alarms, but with significantly greater latency, compared with the moving window method and the time-averaging method. Time averaging showed significantly fewer patients affected by false alarms compared with moving window, and without latency. These findings highlight how there are continual monitoring applications for which the proposed test characteristics provide additional, useful information. Metrics of consistency and latency can demonstrate additional properties that are likely relevant to clinical practice.

REFERENCES

1. Korniewicz DM, Clark T, David Y. A national online survey on the effectiveness of clinical alarms. Am J Crit Care. 2008;17(1): 36–41.

2. Imhoff M, Kuhls S. Alarm algorithms in critical care monitoring. Anesth Analg. 2006;102(5):1525–37.

3. Chambrin MC. Alarms in the intensive care unit: how can the number of false alarms be reduced? Crit Care. 2001;5(4):184–8.

4. Lawless ST. Crying wolf: false alarms in a pediatric intensive care unit. Crit Care Med. 1994;22:981–5.

5. Wald A. Sequential tests of statistical hypotheses. Ann Math Statist. 1945;16(2):117–86.

6. Wald A. Sequential Analysis. New York: John Wiley and Sons; 1947.

7. Reynolds MR, Kim K. Multivariate control charts for monitoring the process mean and variability using sequential sampling. Seq Anal. 2007;26(3):283–315.

8. Bogowicz P, Flores-Mir C, Major PW, Heo G. Sequential analysis applied to clinical trials in dentistry: a systematic review. Evid Based Dent. 2008;9(2):55–60.

9. Rácz A. Comments on the sequential probability ratio testing methods. Ann Nucl Energy. 1996;23(11):919–34.

10. Carlin BP, Chaloner K, Church T, Louis TA, Matts JP. Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. Statistician. 1993;42: 355–67.

11. Brown M. Bayesian detection of changes of a Poisson process monitored at discrete time points where the arrival rates are unknown. Seq Anal. 2008;27(1):68–77.

12. Chen L, McKenna TM, Reisner AT, Gribok A, Reifman J. Decision tool for the early diagnosis of trauma-patient hypovolemia. J Biomed Inform. 2008;41:469–78.

13. Cooke WH, Salinas J, Convertino VA, et al. Heart rate variability and its association with mortality in prehospital trauma patients. J Trauma. 2006;60(2):363–70.

14. Chen L, Reisner AT, Gribok A, Reifman J. Exploration of prehospital vital-sign trends for the prediction of trauma outcomes. Prehosp Emerg Care. 2009;13(3):286–94.

15. Reisner AT, Chen L, McKenna TM, Reifman J. Automatically-computed prehospital severity scores are equivalent to scores based on medic documentation. J Trauma. 2008;65(4):915–23.

16. Yu C, Liu Z, McKenna T, Reisner AT, Reifman J. A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms. J Am Med Inform Assoc. 2006;13(3):309–20.

17. Chen L, McKenna TM, Reisner AT, Reifman J. Algorithms to qualify respiratory data collected during the transport of trauma patients. Physiol Meas. 2006;27(9):797–816.

18. Chen L, Reisner AT, Gribok A, McKenna TM, Reifman J. Can we improve the clinical utility of respiratory rate as a monitored vital sign? Shock. 2009;31(6):574–80.

19. Liddell FDK. Simplified exact analysis of case-referent studies: matched pairs; dichotomous exposure. J Epidemiol Community Health. 1983;37:82–4.

20. Rhee KJ, Willits NH, Turner JE, Ward RE. Trauma Score change during transport: is it predictive of mortality? Am J Emerg Med. 1987;5:353–6.

21. Shapiro NI, Kociszewski C, Harrison T, Chang Y, Wedel SK, Thomas SH. Isolated prehospital hypotension after traumatic injuries: a predictor of mortality? J Emerg Med. 2003;25:175–9.

22. Lipsky AM, Gausche-Hill M, Henneman PL, et al. Prehospital hypotension is a predictor of the need for an emergent, therapeutic operation in trauma patients with normal systolic blood pressure in the emergency department. J Trauma. 2006;61:1228–33.

23. Duckitt RW, Buxton-Thomas R, Walker J, et al. Worthing physiological scoring system: derivation and validation of a physiological early-warning system for medical admissions. An observational, population-based single-centre study. Br J Anaesth. 2007;98(6): 769–74.

24. Subbe CP, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. Emerg Med J. 2006;23(11):841–5.

25. Peres Bota D, Melot C, Lopes Ferreira F, Vincent JL. Infection Probability Score (IPS): A method to help assess the probability of infection in critically ill patients. Crit Care Med. 2003;31(11): 2579–84.

26. Cuthbertson BH. Optimising early warning scoring systems. Resuscitation. 2008;77(2):153–4.

27. Gelinas JF, Davis GM, Arlegui C, Cote A. Prolonged, documented home-monitoring of oxygenation in infants and children. Pediatr Pulmonol. 2008;43(3):288–96.

28. Sibanda N, Lewsey JD, van der Meulen JH, Stringer MD. Continuous monitoring tools for pediatric surgical outcomes: an example using biliary atresia. J Pediatr Surg. 2007;42(11):1919–25.

29. Poloniecki J, Sismanidis C, Bland M, Jones P. Retrospective cohort study of false alarm rates associated with a series of heart operations: the case for hospital mortality monitoring groups. BMJ. 2004;2004:328–75.

30. Reynolds MR, Stoumbos ZG. The SPRT chart for monitoring a proportion. IIE Trans. 1998;30(6):545–61.

31. Musonda P, Hocine MN, Andrews NJ, Tubert-Bitter P, Farrington CP. Monitoring vaccine safety using case series cumulative sum charts. Vaccine. 2008;26(42):5358–67.

32. Sebille V, Bellissant E. Comparison of four sequential methods allowing for early stopping of comparative clinical trials. Clin Sci (Lond). 2000;98(5):569–78.

33. Golhar DY, Pollock SM. Sequential analysis for diagnosing diabetes. Med Decis Making. 1987;7(1):47–51.

34. Griffin MP, O'Shea TM, Bissonette EA, Harrell FE, Jr., Lake DE, Moorman JR. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. Pediatr Res. 2003;53(6):920–6.

35. Sauaia A, Moore FA, Moore EE, Moser KS, Brennan R, Read RA, Pons PT. Epidemiology of trauma deaths: a reassessment. J Trauma. 1995;38:185–93.

36. Peng R, Chang C, Gilmore D, Bongard F. Epidemiology of immediate and early trauma deaths at an urban level I trauma center. Am Surg, 1998;64:950–4.